# User-aware page classification in a search engine

Rachel Aires
NILC/ICMC, University of São Paulo
Caixa Postal 16668 13560-970
São Carlos/SP, Brazil
raires@icmc.usp.br

Sandra Aluísio
NILC/ICMC, University of São Paulo
Caixa Postal 16668 13560-970
São Carlos/SP, Brazil
sandra@icmc.usp.br

Diana Santos
Linguateca, SINTEF ICT
Pb 124 Blindern, 0314
Oslo, Norway
diana.santos@sintef.no

## ABSTRACT
In this paper we investigate the hypothesis that classification of Web pages according to the general user intentions is feasible and useful. As a preliminary study we look into the use of 46 linguistic features to classify texts according to genres and text types; we then employ the same features to train a classifier that decides which possible user need(s) a Web page may satisfy. We also report on experiments for customizing searching systems with the same set of features to train a classifier that helps users discriminate among their specific needs. Finally, we describe some user input that makes us confident on the utility of the approach.

## Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]

## General Terms
Algorithms, Design, Reliability, Experimentation, Human Factors, Languages.

## Keywords
Natural language processing, machine learning, information retrieval, text categorization, web search, stylistic features, personalised search.

## 1. STYLE IN INFORMATION ACCESS
One of the goals of the present workshop is to discuss, among others, the question: Can stylistic information be used profitably e.g. in information access interfaces? We believe our work in text categorization for IR shows that this is definitely the case, although a consensual definition of style is a pre-requisite to agreement on this matter.

For us, stylistic information is that part of linguistic data that is not related to content, but rather with the way the content is conveyed. Content and style are, however, intricately mixed and are related in complex ways. While some style features are individual and often involuntary, others are taught and described in style manuals and professional training. Style is a set of elusive properties that allow scholars to talk about genres, and humans in general to assess what is appropriate or not in specific (con)texts.

Style, as most linguistic concepts, can be approached at least from two angles: as a macro property of full texts (and/or collections of texts), something that can only be predicated of a large collection, or as a micro property that is operational in every minor linguistic choice a speaker or writer makes. The challenge in automatic style categorization is to connect the two, and, using easy to compute features, provide a classification in terms of recognizable kinds.

The kinds of texts we looked into, or rather the classification we wanted to get at, was not of style in itself, but of what kind of user's need a particular Web page was supposed to satisfy. There were two reasons for this somewhat radical move:

1) automatic style categorization is generally motivated by the belief that style is a good indicator of different user needs, as in Kesser et al. (1997:32): "in information retrieval, genre classification could enable users to sort search results according to their immediate interests" [1]. Why go indirectly through genre and not directly to user's interests?

2) there was not a well-established off-the-shelf classification scheme for the Web, with some people claiming there are many new and evolving genres in it. For example, [2] and [3] Web typologies are significantly different.

Also, genre and style are often connected with the text producers, while our view was a consumer's view. Given the well-know mismatch between what users want and what producers offer, it is not obvious that one should start by looking at the collection instead of looking at what the presumable goals of users might be. Inspired by Broder [4] and previous work in detecting user's goals in Web search [5], we devised a user need typology from a qualitative analysis of the TodoBr logs.

Our assumption is that categorizing the results by user need (i.e., which would user need they would satisfy) would improve information access, something related, but not necessarily equivalent to what was reported and tested by Bretan et al [2].

Currently, users are faced with information overflow: the problem is too large quantity, not scarcity of information. Information Retrieval is now advancing towards different ways of organizing information. Several search engine companies have recently announced initiatives to improve search results by allowing users to customize their searches by delimiting the search context, in several ways. We are currently aware of 1) the saving of previous queries and more general user behaviour tracking; 2) the ability for the user to define subject profiles of interest, and 3) the offer of vertical search within the web, as in geographically-aware search, news lookup and search for famous people. All of these developments have, in addition to privacy issues, a particular problem, namely how to maintain profile accuracy across different tasks and over extended periods, something arguably difficult to deal with. In fact, the user's focus and interests can

change, and repeated frequent searches may be too specific, as in tasks from the user's work. Cognitive aspects of search behaviour have been claimed to deserve close attention, especially because multitasking information seeking seems to be common in the Web as well as in other information seeking environments [6].

So, in [7] we took the path to present to the user the results of her search classified by type of goal, i.e., develop a categorization meant to improve the presentation of the search results. Eventually this kind of categorization might also be employed for indexing, but this has not been the focus of our work.

Our work was inspired by Karlgren's [8] studies of systematic stylistic variation in order to characterize the genre of documents and improve Web search. Later on, we performed also similar experiments on automatic genre classification, but using a different genre classification, developed for Brazilian Portuguese in connection with the freely available reference corpus Lácio-Ref [9].

Our primary goal is to find which stylistic features can be used to classify web pages in Portuguese into understandable and useful classes to the web search task, in order to decrease the user effort to retrieve information.

In this paper, we report on several experiments performed to investigate automatic web pages classification into:

- genre and text types;
- seven general kinds of user needs;
- personalized user needs.

The paper is structured as follows: we start by presenting our explorations in genre and text type classification. Then, we present new experiments carried out to classify texts according to seven users' needs. Finally, we present two studies on the use of stylistic features to create customized classification schemes, one of which for English. We end the paper with a discussion of these results and alleys for further work.

## 2. CLASSIFICATION IN GENRES AND TEXT TYPES

The genre of a text captures its communicative intention and discourse character. In other words, it classifies the community to which the text is addressed and the human activities that make it relevant. Genres can be told apart by the text types (which are defined by a particular text structure, lexicon, syntax, and adequacy to the main theme) usually associated to each of them.

Karlgren, based primarily in Biber [10], used stable characteristics of texts for genre categorization. According to Karlgren [8], style is the difference between two ways of saying the same thing, and systematic stylistic variation can be used to characterize the genre of documents. In one of his studies ([8]: Chapter 16) he looked into the design of an interactive system with the interface incorporating stylistic information, categorizing retrieval results by genre, and displaying the results using this categorization. In this experiment, eleven categories were employed and a user-centred evaluation was performed. The users were asked to execute two tasks each, using the interface prototype with stylistic features and the web search engine Altavista. Karlgren concluded that most users used the interface as intended and many searched for documents in the genres the results could be expected to show up in.

Biber [10] has studied English text variation using several variables, and found that texts vary along five dimensions. Registers would then differ systematically along each of these dimensions, relating to functional considerations such as interactiveness, involvement, purpose, and production circumstances, all of which have marked correlates in linguistic structure.

Stamatatos et al [3] have also worked with genre classification based on stylometric methods, creating a Web corpus for Modern Greek and automatically categorizing it.

We performed the following experiment: We used the genres scheme presented by Aluísio et al [9] on which the Lácio-Ref[1] corpus was based, in connection with the corpus for training. The corpus has 4,278 files with 8,291,818 words, divided into 5 genres (scientific, informative, law, literary, and instructional) and 30 text types (paper, administrative circular, statement, dissertation, editorial, interview, law, textbook, public notice, decree, short stories, letter, monograph, news, legal opinion, report, review, abstract, provisional measure, official letter, ordinance, receipt, news reporting, resolution, government body rules, court management measure, court decision, superior court decision, poem, and other).

To make text classification even more flexible an option is to allow the user to get the search results classified by text types. For example, it is improbable that the same information need can simultaneously be satisfied with a poem about lonely hearts and a recipe using chicken heart as an ingredient. Therefore, we investigated whether classification of Web pages into text types could mirror somehow user's intentions.

We computed the 46 features for each text that had been suggested in [7] (shown in Figure 1), and used them to train a genre classifier. These features, which are mainly closed lists, were inspired by those proposed by Biber [10] and Karlgren [8], but checked in grammars and textbooks for Portuguese.

| **Word-based statistics** |
|---|
| Type/token ratio |
| capital type token ratio |
| digit content |
| average word length in characters |
| long words (>6 chars) count |
| **Text-based statistics** |
| Character count |
| average sentence length in characters |
| sentence count |
| average sentence length in words |
| text length in words |
| **Other statistics** |
| the subjective markers "acho", "acredito que", "parece que" and "tenho impressão que" ("I think so", "I believe that", "it seems that", "have the impression that") |

[1] http://www.nilc.icmc.usp.br/lacioweb/

| |
|---|
| the present forms of verb to be "é/são" ("is/are") |
| the word "que" (can be: noun, pronoun, adverb, preposition, conjunction, interjection, emphatic particle) |
| the word "se" ("if/whether" and reflexive pronoun) |
| the discourse markers "agora", "da mesma forma", "de qualquer forma", "de qualquer maneira" and "desse modo" ("now", "on the same way", "anyway", "somehow" and "this way") |
| the words "aonde", "como", "onde", "por que", "qual", "quando", "que" and "quem" on the beggining of questions (wh-questions) |
| "e", "ou" and "mas" as sentence-initial conjunctions ("and", "or", "but") |
| amplifiers. Amplifiers scale upwards (Quirk et al, 1992), denoting either an upper extreme of a scale or a high degree, high point on the scale. Some examples are: "absolutamente" (absolutely), "extremamente" (extremely), "completamente" (completely) and "longe" (far). |
| conjuncts. Most conjuncts are adverbs and prepositional phrases (Quirk et al, 1992). Some examples are: "além disso" (moreover), "conseqüentemente" (accordingly), "assim" (thus) and "entretanto" (however). |
| downtoners. Downtoners have a lowering effect on the force of the verb and many of them scale gradable verbs, they can have a slight lowering effect, scale downwards considerably or serve to express an approximation to the force of the verb (while indicating its non-application) (Quirk et al, 1992). Some examples are: "com exceção" (with the exception), "levemente" (slightly), "parcialmente" (partially) and "praticamente" (practically). |
| emphatics. Emphatics (emphasizers) have a general heightening effect (Quirk et al, 1992). Some examples are: "definitivamente" (definitely), "é óbvio que" (it is obvious that), "francamente" (frankly) and "literalmente" (literally). |
| suasive verbs. Some examples are the verbs: *aderir* (to adhere), *distinguir* (to distinguish), *crer* (to believe) and *dar* (to give). |
| private verbs. Some examples are the verbs: *partir* (to leave), *ter* (to have), *averiguar* (to check) and *guardar* (to keep). |
| public verbs. Some examples are the verbs: *abolir* (to abolish), *promulgar* (to promulgate), *mencionar* (to mention) and *declarar* (to declare). |
| number of definite articles |
| number of indefinite articles |
| first person pronouns |
| second person pronouns |
| third person pronouns |
| number of demonstrative pronouns |
| indefinite pronouns and pronominal expressions |
| number of prepositions |
| place adverbials |
| time adverbials |
| number of adverbs |
| number of interjections |
| contractions |
| Causative conjunctions |
| Final conjunctions |
| Proportional conjunctions |
| Temporal conjunctions |
| Concessive conjunctions |
| Conditional conjunctions |
| "conformative" conjunctions |
| comparative conjunctions |
| consecutive conjunctions |

**Figure 1. The 46 features selected**

We used the Weka J48, Sequential Minimal Optimization (SMO) and Logistic Model Tree (LMT) algorithms [11]. J48 is the Weka implementation of the decision tree learner C4.5. C4.5 was chosen for several reasons: it is a well-known classification algorithm, it had already been used in similar studies [8], and it produces easily understandable rules. LMT [12] is a classification algorithm for building 'logistic model trees', which are classification trees with logistic regression functions at the leaves. SMO implements Platt's [13] sequential minimal optimisation algorithm for training a support vector classifier using scaled polynomial kernels, transforming the output of SVM into probabilities by applying a standard sigmoid function that is not fitted to the data. The implementation used does not perform speed-up for linear feature space and sparse input data. It globally replaces all missing values, transforms nominal attributes into binary ones, and normalizes all numeric attributes.

The results[2] for precision, recall and F-measure are shown in Table 1.

**Table 1. Results for genres and text types**

| ***Algorithms*** | **J48** | **SMO** | **LMT** |
|---|---|---|---|
| **Classification in Genres** | | | |
| Precision | 0.82 | 0.81 | 0.89 |
| Recall | 0.77 | 0.85 | 0.85 |
| F-measure | 0.79 | 0.82 | 0.87 |
| **Classification in Text Types** | | | |
| Precision | 0.65 | 0.55 | 0.76 |
| Recall | 0.67 | 0.91 | 0.74 |
| F-measure | 0.65 | 0.69 | 0.75 |

Results for genres confirm that stylistic features can also be used in the classification of Portuguese texts, as it was done in studies for English and other languages. The best result was achieved with LMT. The results for text types were poorer, even for the

[2] In all experiments presented in this paper we used 10-fold cross validation.

best algorithm. Reasons for this may include the fact that the corpus is not balanced in terms of text types (there are 300 texts for some types, while for others there are only 6), or that the text types themselves do not really stand apart in linguistic terms.

## 3. CLASSIFICATION IN SEVEN USERS' NEEDS

As documented in [7], the classification scheme based on the seven users' needs was the outcome of a qualitative analysis of the most common users' needs for the period between November 1999 and July 2002, provided by TodoBr[3] logs — a major Brazilian search engine from Akwan Information Technologies. This classification reflects what the user wants:

1) A definition of something or to learn how or why something happens. For this need, dictionaries, encyclopaedias, textbooks, technical articles, reports and texts of the informative genre would present the best results.

2) To learn how to do something or how something is usually done, as in finding a recipe of cake or learning to make gift boxes and installing Linux. Typical results are texts of the instructional genre, such as manuals, textbooks, readers, recipes and even some technical articles or reports.

3) A comprehensive presentation about a given topic. In this case, the best results should be texts of the instructional, informative and scientific genres, e.g. textbooks, essays and long articles.

4) To read news about a specific subject, as the news about the current situation in a given part of the world, or the latest results of soccer games. The best answers in this case would be texts of the informative genre, e.g. online newspapers and magazines.

5) To find information about someone or a company or organization. A typical example would be the user interested in more information about his/her blind date or to find the contact information of someone he met in a conference. Typical answers here are personal, corporation and institutional web pages.

6) To find a specific web page whose URL the user does not remember. For this type of need the results could be from any type of text or genre. The only way to identify this need would be if the interface asked the user what type of page he/she is looking for.

7) To find URLs where for accessing online services, such as buying clothes or downloading software. The best answer to this kind of request is commercial text types (companies or individuals offering products or services).

In a previous experiment (see [7] for more details) we created a corpus with 511 texts extracted from the Web, 73 for each type of need[4] plus additional 73 texts that would not answer any of the six types used (we call it "others"), in order to have a balanced corpus. The resulting corpus had 640,630 words. For comparison, note that Biber's 481 texts amounted to a corpus with approximately 960,000 words, which is larger in number of words because Web texts tend to be smaller.

[3] www.todobr.com.br

[4] Except for type 6, which, as explained above, can correspond to any kind of text.

For this experiment we used the 46 features shown in Figure 1. We computed these statistics for all texts, and trained classifiers using 2, 3 (2 categories plus "others"), 4, 5 (4 categories plus "others"), 6 and 7 categories (6 categories plus "others") (Table 2). In [14] we used mainly the J48 algorithm.

**Table 2. Used categories**

| 2 categories | 4 categories | 6 categories |
|---|---|---|
| 1) the union of needs 1, 2, 3, 4 and 5<br>2) need 7 | 1) the union of needs 1, 2, 3<br>2) need 4<br>3) need 5<br>4) need 7 | Need 1<br>Need 2<br>Need 3<br>Need 4<br>Need 5<br>Need 7 |

The classification with 2 categories decides whether a page gives any kind of information about a topic or gives access to an online service. The classification with 4 categories distinguishes among information about something, someone or some company/institution/organization, news, and online services. Finally, the classification with 6 categories is the most comprehensive presented here, which excludes only category 6 that can be of any type of text or genre. The class "others" contains text types like blogs, jokes, poetry, etc, that are examples of text types not covered by the seven users' needs.

The results were encouraging: We got 90.93% of correct classification for 2 categories, 76.97% for 3, 65.06% for 4; 56.56% for 5; 52.01% for 6 and 45.32% for 7 categories. We then replicated these experiments using all 44 Weka algorithms which could deal with non-nominal features, with non-numerical classes, with the number of classes we needed (maximum 7) and which did not present errors related to the standard deviation of our features for any of our classes. Fourteen algorithms achieved the same or better results than J48, regarding the percentage of correct decisions. The best ones were LMT and SMO. The best result for 2, 3, 4, 5, 6 and 7 categories were, respectively 93.83%, 82.97%, 73.74; 67.90%, 63.69% and 58.31% (see [14] for a full description of the results, such as precision and recall per class).

In spite of these good results, there were problems in this approach, particularly the assumption that any given text could only satisfy one user's need. So we created a new and larger corpus, "Yes, user!",[5] which was reclassified in as many of 22 classes (see [15] for corpus description), some with only a few texts. In order to have a balanced corpus, it was enlarged to 1,703 texts (2,159,491 words).

We carried out 3 experiments using the reclassified corpus with: (i) the 46 features of Figure 1; (ii) the 46 features from (i) plus 5 functions to measure vocabulary richness, taken from [3] and shown in Figure 2, resulting in 51 features; (iii) the features from (i) plus features dealing with the most frequent words in the corpus, after eliminating stop-words, linking verbs, adverbs, domain related words (terminology) and further grouping some words together (108 features).

[5] http://www.linguateca.pt/Repositorio/YesUser/

In Figure 2 $V_i$ is the number of words used exactly i times and α is fixed as 0.17.

$$K = \frac{10^4 (\sum_{i=1}^{m} i^2 V_i - N)}{N^2} \qquad W = N^{V^{-\alpha}}$$

$$S = \frac{V_2}{V}$$

$$R = \frac{(100 \log N)}{(1 - (\frac{V_1}{V}))} \qquad D = \sum_{i=1}^{V} V_i \frac{i(i-1)}{N(N-1)}$$

**Figure 2. functions to measure vocabulary richness**

In the first experiment we generated 3 classification schemes[6]: one with all the six needs, another distinguishing among pages which offer services, pages which offer information and pages which offer both, and the last one which distinguishes between services and information. The results are shown in Table 3.

**Table 3. Correct classifications using the 46 features**

| | J48 | SMO |
|---|---|---|
| Full classification in 6 needs plus "others" | 69.7% | 72.52% |
| Information x service x information and service plus "others" | 72.17% | 73.58% |
| Information x service plus "others" | 85.11% | 86.37% |

The second and the third experiments were done only with the full classification scheme (six needs plus "others") and the results are shown in Table 4.

**Table 4. Correct classifications using 51 or 108 features**

| | J48 | SMO |
|---|---|---|
| **51 features** | 70.38% | 73.77% |
| **108 features** | 73.17% | 77.02% |

The results from table 3 and 4 are significantly better than those in [10] which presented a precision of 45.32% for the classification in six categories plus "others" and 82.97% for the classification in two categories plus "others". For 6 categories plus "others" the best result was with 108 features and SMO; for 2 categories plus "others" the best result was with SMO.

# 4. CREATION OF CUSTOMIZED CLASSIFICATION SCHEMES

Obviously, the seven types of user needs explained in Section 3 do not cover all kinds of user intentions, as users may do all kinds of unpredictable searches and it is unlikely that one can recover their intentions by looking only at the logs. However, the very features used to generate rules and classify texts can be used to build customized schemes for other tasks. For example, a doctor can create a classification scheme to distinguish between web pages with technical articles about a disease and web pages that deal with the subject without scientific rigor. However, it is not possible to use the same features we have studied to distinguish among subjects, for example, to tell cardiology technical texts apart from other medical technical texts. We plan to offer customized schemes to the user in a desktop web search prototype being developed, where the user can select examples of text types that often make his/her searches difficult. In the doctor's example, he/she would give to the system samples of technical and non-technical material that would be used as training material. The system would then automatically calculate the features for the given text set, train a classifier and present an estimation of the system efficacy to the user personal scheme. The generated classification model would be saved as a new option of the classification task. Summing up, we would offer predefined options (genres, text types and seven user's needs) as it is provided by search engines shortcuts and tabs, but we will also allow the user to create his/her own shortcut specific to the binary text type related problematic tasks that he/she often performs.

In the following sections we show three case studies regarding the use of stylistic features to create customized classification schemes.

## 4.1 Legal texts

We created a corpus with 200 texts in the law domain, extracted from the Web. Half of them are meant for experts, the other half for laymen. In order to find out how many texts are necessary for training personalized schemes, in this experiment we have used: (i) an increasing number of texts in the training sets; (ii) the algorithms J48, SMO and LMT; (iii) the 46 features from Figure 1. Results of each classifier appear in Table 5.

**Table 5. Results for legal texts in Portuguese**

| | J48 | SMO | LMT |
|---|---|---|---|
| **20 texts** | | | |
| **Precision** | 0.43 | 0.61 | 0.42 |
| **Recall** | 0.60 | 0.79 | 0.56 |
| **F-measure** | 0.48 | 0.67 | 0.47 |
| **100 texts** | | | |
| **Precision** | 0.67 | 0.78 | 0.81 |
| **Recall** | 0.68 | 0.75 | 0.75 |
| **F-measure** | 0.66 | 0.75 | 0.76 |
| **200 texts** | | | |
| **Precision** | 0.77 | 0.83 | 0.84 |
| **Recall** | 0.76 | 0.84 | 0.84 |
| **F-measure** | 0.76 | 0.83 | 0.83 |

The best results were achieved with a training set with 200 texts and the algorithms SMO and LMT.

We have also trained a classification scheme for texts in English using: (i) a corpus with 200 texts extracted from www.findlaw.com; (ii) the algorithms J48, SMO and LMT and (iii) 52 features taken from Biber and Karlgren [1, 2] which are the original features for English that were adapted for Portuguese (Figure 1) plus 3 types of modals, 2 of negation, nominalizations, besides reflexive and possessive pronouns. Results of each

[6] In all three schemes the class "others" was considered.

classifier appear in Table 6. Figure 3 shows the J48 decision tree when trained with the 200 texts.

**Table 6. Results for legal texts in English**

| | J48 | SMO | LMT |
|---|---|---|---|
| **20 texts** | | | |
| **Precision** | 0.82 | 0.78 | 0.77 |
| **Recall** | 0.88 | 0.78 | 0.80 |
| **F-measure** | 0.84 | 0.77 | 0.78 |
| **100 texts** | | | |
| **Precision** | 0.87 | 0.95 | 0.91 |
| **Recall** | 0.86 | 0.91 | 0.86 |
| **F-measure** | 0.85 | 0.92 | 0.87 |
| **200 texts** | | | |
| **Precision** | 0.89 | 0.96 | 0.94 |
| **Recall** | 0.87 | 0.92 | 0.92 |
| **F-measure** | 0.87 | 0.94 | 0.93 |

The best results were achieved with a training set with 200 texts and the algorithm SMO.

```
second person pronoun <= 0.062305
|  capital type token ratio <= 106
|  |  predictive modals <= 0.295683: laymen (4.0)
|  |  predictive modals > 0.295683: expert (2.0)
|  capital type token ratio > 106
|  |  second person pronoun <= 0.022763: expert (82.0)
|  |  second person pronoun > 0.022763
|  |  |  definite article <= 3.540519: laymen(4.0/1.0)
|  |  |  definite article > 3.540519: expert (7.0)
second person pronoun > 0.062305
|  interjections <= 0.006979
|  |  prepositions <= 8.235294: laymen(92.0/1.0)
|  |  prepositions > 8.235294
|  |  |  second person pronoun <= 0.160128: expert (17.0/1.0)
|  |  |  second person pronoun > 0.160128
|  |  |  |  prepositions <= 11.081323
|  |  |  |  |  definite article <= 5.350978
|  |  |  |  |  |  synthetic negation <= 0.26178: laymen(48.0/4.0)
|  |  |  |  |  |  synthetic negation > 0.26178: expert (3.0/1.0)
|  |  |  |  |  definite article > 5.350978
|  |  |  |  |  |  type token ratio <= 0.357788: expert (8.0)
|  |  |  |  |  |  type token ratio > 0.357788: laymen(2.0)
|  |  |  |  prepositions > 11.081323
|  |  |  |  |  first person pronouns <= 0.151976: expert (9.0)
|  |  |  |  |  first person pronouns > 0.151976: laymen(2.0)
|  interjections > 0.006979
|  |  prepositions <= 4.71464: laymen(2.0)
|  |  prepositions > 4.71464: expert (17.0)
```

**Figure 3. Decision tree for English texts classification scheme**

## 4.2 Finding product descriptions

The second study concerned finding out whether E-commerce pages described products on sale or not. We used a collection provided by Martins & Moreira [16] containing 1,252 pages.

**Table 7. Results for e-commerce pages**

| | J48 | SMO | LMT |
|---|---|---|---|
| **Precision** | 0.87 | 0.90 | 0.90 |
| **Recall** | 0.85 | 0.69 | 0.86 |
| **F-measure** | 0.86 | 0.78 | 0.88 |

The best results were achieved with LMT.

# 5. EVALUATING THE SCHEMES

In order to have some feedback from potential users, we applied a questionnaire to undergraduate students of computer science, linguistics, medicine and to graduate photography students. The goals were to find out:

- How clear to the users was the seven user's needs scheme
- How clear was the genre classification scheme. This was done in 2 ways: (i) asking if any of the three genre schemes presented in [9, 8:16, 3] was helpful for the search task; (ii) presenting the genre scheme from [9] through text type examples and calling it text types schemes (we did not present it as 30 classes mentioned in Section 2, we presented it in 9 classes)
- Which schemes were easier to use
- Whether the user would spend one day collecting text samples to generate a system that would be specific for the tasks that often trouble him

Sixty three students answered the questionnaire. At least two students believed that one of the schemes above was not helpful, specifically: 2 for the seven user's needs, 3 for the text types, 8 for the genre scheme presented in [8:16], 12 for the genre scheme presented in [9] and 13 for the genre scheme presented in [3]. At least six students believed that one of the schemes was easier to use: 25 for the seven user's needs, 29 for the text types, 15 for the genre scheme presented in [8:16], 13 for the genre scheme presented in [9] and 6 for the genre scheme presented in [3].

The hypothesis behind our work was that it is easier for a user to choose among types of needs than between genres. From the questionnaire we realized that the students did not completely understand the genres labels, since the difference between the genres scheme and the text types scheme was only the label and only 3 considered it not useful while 12 considered the genres scheme one not useful. As an example for the labelling, the label for the instructional genre was changed to "text book, culinary recipe, course notes, etc."

The number of students which considered the scheme based on the seven user's needs useful was also larger than those that preferred the genre scheme. However this has to be confirmed using a user-centred evaluation of our prototype. Figure 4 shows a screen dump of our desktop meta searcher prototype, named Leva-e-Traz ("takes and brings").

The results seem to indicate that there is something to be gained classifying Web pages using the a priori schemes: seven users' needs, genres scheme and text type scheme (the one presented in Section 2). All 41 users who had reported having frequent problems in their searches answered that they would spend a day

creating personalised schemes, which apparently confirms the feasibility of the option described in Section 4.


**Figure 4. Leva-e-Traz main screen**

## 6. DISCUSSION AND ONGOING WORK

In this paper we have presented results for genre, text types, seven user needs and personalized classification of texts on the Web.

On the one hand, we confirmed that the use of stylistic features to classify texts in genre and text types, as advocated and used for other languages, also works for Portuguese.

In addition, we believe that our attempt to automatically categorise, in terms of user needs, texts on the Web – first reported in [7] – had not been tried before, for any language. We applied it to Brazilian Portuguese, and the experiments reported here improved precision from 0.45 reported in [7] to 0.77, which seems to indicate that this 7 types can be reliably enough identified to help the user.

Finally, we have also obtained some first results for personalized classification, achieving a precision of at least 0.84. As far as we know, the decision of how to classify and rank the texts has not been previously put in the user's hands, although adaptive systems learning from user choices exist in the literature [17]. We are currently conducting more experiments like the ones presented in Section 4 to find out how many texts the user has to collect, and how the selection can be improved. If we confirm, with future experiments, that a small number of texts, such as 200, is sufficient to achieve good results, we may have found a cost-effective way to solve a user's specific text type related problem, as well as sharpened our knowledge of which relevant features to add.

We are also currently investigating the addition of structural clues (such as those in HTML) and of more deep linguistic features (such as those provided by syntactically parsing the text) to our classifiers, and hope to report on these experiments soon [18].

## 7. ACKNOWLEDGEMENTS

Our thanks to Akwan Information Technologies for the TodoBr logs. This work was partially supported by grant POSI/PLP/43931/2001 from Fundação para a Ciência e Tecnologia (Portugal), co-financed by POSI.

## 8. REFERENCES

[1] Kessler, B.; Nunber, G.; Schütze, H. Automatic Detection of Text Genre, in Proceedings of the 35th annual meeting on Association for Computational Linguistics (Morristown, NJ, USA, 1997), ACL, 32-38.

[2] Bretan, I.; Dewe, J.; Hallberg, A.; Wolkert, N.; Karlgren, J.: Web-Specific Genre Visualization, *WebNet '98* (Orlando, Florida, November 1998).

[3] Stamatatos, E.; Kakotakis, N.; Kokkinakis, G. Automatic text categorization in terms of genre and author. Computational linguistics (2001), vol 26, number 4, 271-295.

[4] Broder, A. "A Taxonomy of Web Search", SIGIR Forum 36 (2), Fall 2002, p.3-10.

[5] Aires, R.; Aluísio S.: Como incrementar a qualidade das máquinas de busca: da análise de logs à interação em Português. *Revista Ciência da Informação*, **32** (1), 2003, pp. 5-16.

[6] Spink, A.; Ozmutlu , H. C.; Ozmutlu, S.: Multitasking information seeking and searching processes. *Journal of the American Society for Information Science and Technology* **53** (8), 2002, pp. 639-652.

[7] Aires, R.; Manfrin, A.; Aluísio, S.; Santos, D. (2004) What is my Style? Using Stylistic Features of Portuguese Web Texts to classify Web pages according to Users' Needs. In LREC 2004. Lisbon - Portugal, May 2004, p. 1943-1946.

[8] Karlgren, J.: Stylistic Experiments for Information Retrieval. PhD Thesis, Stockholm University, Department of linguistics, 2000.

[9] Aluisio, S.; Pinheiro, G.; Finger, M.; Nunes, M.G.V.; Tagnin, S.E. The Lacio-Web Project: overview and issues in Brazilian Portuguese corpora creation. In Proceedings of Corpus Linguistics (2003), Lancaster, UK. v. 16, p. 14-21.

[10] Biber, D.: Variation across speech and writing. Cambridge University Press. Cambridge, UK (1988)

[11] Witten, I. H., Frank, E.: Data Mining: Practical machine learning tools with Java implementations. San Francisco: Morgan Kaufmann, 2000.

[12] Landwehr, N, Hall, M, Frank, E. (2003) Logistic Model Trees. ECML 2003, p. 241-252.

[13] Platt, J.: Fast training of support vector machines using sequential minimal optimization. In Advances in kernel methods: support vector learning. B. Schölkopf, C. Burges, and A. Smola, eds. MIT Press, 1999, p. 185-208.

[14] Aires, R.; Manfrin, A.; Aluísio, S.; Santos, D.: Which classification algorithm works best with stylistic features of Portuguese in order to classify web texts according to users' needs? Relatório técnico nº 241, outubro de 2004, ICMC/USP.

[15] Aires, R.; Aluísio, S.; Santos, D.: "Yes, user!": compiling a corpus according to what the user wants. Proceedings of Corpus Linguistics 2005 (Birmingham, UK, July 14-17 2005).

[16] Martins Junior, J.; Moreira, E. S. Using Support Vector Machines to Recognize Products in E-commerce Pages. In Proceedings of The IASTED International Conference, February 2004, p. 212-217.

[17] Ciravegna, F.; Wilks, Y.: Designing Adaptive Information Extraction for the Semantic Web in Amilcare, in S. Handschuh and S. Staab (eds), *Annotation for the Semantic Web*, Amsterdam: IOS Press, 2003.

[18] Aires, R.: O uso de características lingüísticas para a apresentação dos resultados de busca na Web de acordo com a intenção da busca do usuário – uma instanciação para o português. PhD Dissertation, Computer Science Department (ICMC), University of São Paulo, forthcoming.